Stack Height Analysis for FinFET Logic and Circuit

Xinfei Guo and Qing Qin
Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia
{xg2dt, qq3za@virginia.edu}

***Abstract*— FinFET has appeared to be a good candidate for further extending the technology to the nanoscale regime due to its excellent electrostatic properties and comparative ease of fabrication. Due to little/no body effect in FinFET devices, there is a good chance of increasing the stack height to enable more complex gates and reduce logic depth, which could further improve metrics, like performance. This paper analyzes the tradeoffs between stack height (fan-in) and area/performance through simulating several logic gates based on PTM model. A 64-bit Kogge-Stone adder with different radix size is designed and simulated across five different technology nodes as a case study. The simulation result shows that a radix-4 adder can achieve the best performance, but with area penalty. A design flow that considers all the tradeoffs is also proposed.**

Keywords—FinFET; stack height; body effect; adder; radix

I. INTRODUCTION

With extreme technology scaling, FinFET (or Tri-gate) devices deliver the superior levels of scalability and higher performance and relatively ease of fabrication [1]. FinFET is a promising candidate to be integrated with either bulk technology or silicon-on-insulator (SOI) substrates [2]. It has been lots of study that compare the two [2-3], but it has been demonstrated that SOI FinFETs can offer better electrical characteristic [3], less self-heating problems [3] and less leakage [2]. So FinFET has been mainly fabricated on silicon-on-insulator (SOI) substrates [4]-[5]. The floating body regions of SOI devices are unique because of the box layer separating the transistor devices from the substrate or bulk wafer material. When the box is thick, the back plane potential has very little effect on the channel, and this will make the device threshold insensitive to back plane biasing voltage. Also, since the gate has full control over the channel in FinFET devices, and the body potential will have less effect [6]. All of these made FinFET less affected by body effect, in some FinFET modeling work [7], body effect is totally eliminated.

In some logic cells, NAND for example, several transistors are connected in series and stacked. In planar CMOS circuit, stack height is limited by the body effect. Shown in Fig. 1 is a 4-input NAND cell which includes a 4-stack pull down network. Due to the body effect, voltage between source (Vx) and body (gnd) of the top stacked transistor will increase the threshold voltage and will lead to performance degradation, if the stack height keeps increasing, the pull down current will be very small and the circuit can't even function correctly. For FinFET logic and circuit, due to the insensitivity to the body effect as discussed above, the stack effect will be gone and this will lead to higher stack logic cells with potential of reducing logic depth, and further reducing delay. This will enable several research questions – one is that how to determine the stack height and what is the optimal height in terms of all the metrics; second is that how to design circuit so that it can fully take advantage of this unique property of FinFET. To answer these questions, this paper start from several basic logic gates and simulate based on PTM models. A 64 bit adder designed with different stack height shows the design tradeoffs.

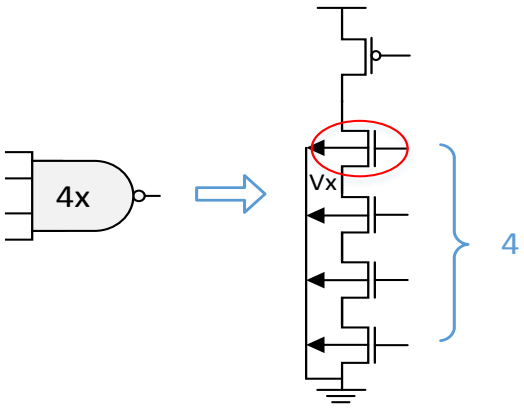


Fig.1. 4-NAND gate with 4 Stacked NMOS

There are a few work looking into similar questions, [8] analyzed the stack sizing of FinFET logic cells in the sub/near-threshold region based on their FinFET models. Their results show that in sub/near-threshold region, the stack depth of 2 is actually highly preferred for FinFET circuit designs. [9] proposed a noise immune leakage tolerant FinFET based wide fan-in dynamic OR gate, but it more focused on noise immune aspects. This work differs from the previous work in the following,

- The simulation and analysis consider the physical aspects of the circuit, wire capacitance for example;
- All simulations are run in super-threshold mode that enables high performance;
- Detailed design tradeoffs are analyzed across technology nodes, and a design flow that balances these tradeoffs is proposed.

The rest of the paper is organized in the following. In Section II, we discuss the FinFET simulation flow. Section III will provide the analysis of some logic cells. In Section IV, we provide a design case to further examine the design tradeoffs. Section V will discuss the results, and a design flow will be also proposed in this section. Section VI will conclude the paper.

II. Simulation Flow

This section will provide details of how we design and simulate FinFET in cadence environment.

A. *FinFET FreePDK15*

FinFET FreePDK15 is an open source 16/20nm FinFET process design kit developed by NCSU [10]. Nangate developed the corresponding open cell library [11]. This will give us a physical aspect of how FinFET logic cell look like. Fig. 2 shows an example of inverter and flip-flop. Based on the layout, one observation is that the P/N ratio is 1:1, thus the FinFET logic cells are more symmetric than the planar CMOS.

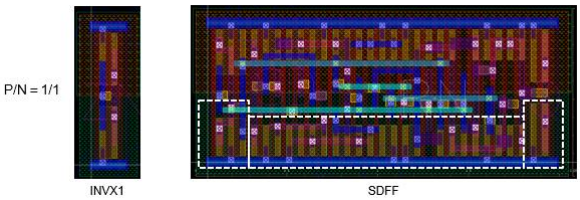


Fig.2. Example layout of FinFET cells

B. *FinFET Models*

The predictive FinFET model, PTM-MG [8] that is based on the BSIM-MG model [13] has been used as the simulation model in this paper. The model includes 5 technology nodes and can be used in cadence *spectre* environment. The simulation flow employed in this paper is shown in Fig. 3. The first step is to create the device symbol that could be used in cadence schematic capture, the symbol in this work is a modified version of FreePDK45. The model still has four terminal with one node as floating. In the actual design, this node can be left open or connected to other terminals. After the symbol is created, some parameter corresponding to FinFET specifically need to be modified. For example, in FinFET, the way to size the transistors is to increase or decrease the number of fingers. So a parameter m is utilized in this paper to adjust the sizing. The last step is to design the circuit and run the simulation by including the PTM model in a similar way to other design kits.

Fig.3. Setup Flow for PTM-MG model

III. Stack Height Analysis

After simulation flow is set up, this section will focus on the stack height analysis in several logic gates and cells.

A. *Body Effect*

To check if the body effect actually affects the circuit, 4-input NAND, 16-input NAND and regular inverter are simulated as shown in Fig. 4. Fig. 5 shows that the average pull-down current for three cases are almost the same. This indicates that the FinFET logic is really insensitive to stack effect and body effect.

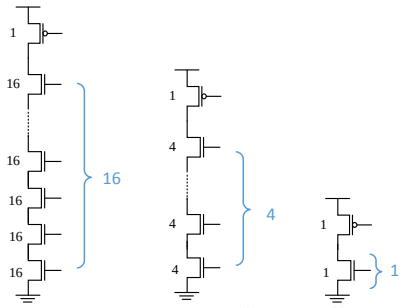


Fig.4. Simulation setup for body effect

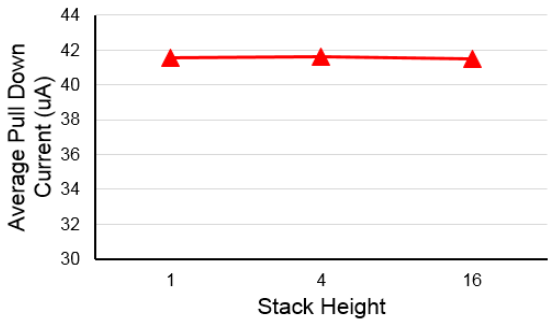


Fig.5. Average pull-down current for different stack height

B. *P/N Ratio*

In last section, the P/N ratio of 1:1 is observed in layout cells. To further investigate if this holds in actual digital design that prefers the balance between PMOS and NMOS, an inverter with P/N of 1:1 is simulated in 16nm HP node. Fig. 6 shows the voltage transfer curve (VTC) under different supply voltages, it shows that the curve is very balanced in all cases. In terms of pull up and pull down delay, the results show that tplh=49.9ps and tphl=45.16ps, which suggest that the ratio of 1:1 is optimal for FinFET design. In this paper, all the cells are sized such that both the pull-up network and the pull-down network have the same driving strength as the 1:1 inverter.

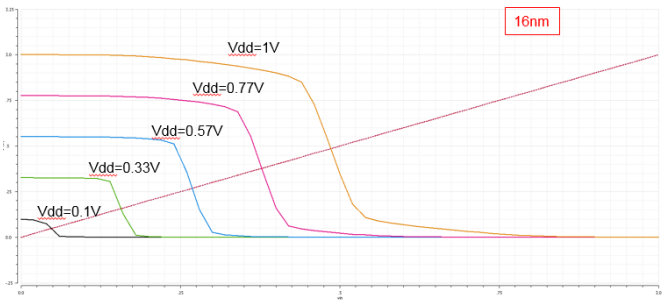


Fig.6. VTC curves under different supply voltages

C. *16-input AND Gate*

To investigate how the stack height will affect the performance, a 16-input AND Gate implemented in 3 different stack height is designed and simulated in 16nm HP node. Fig. 7 shows the schematics of the design, all designs drive a load capacitor of 5fF. The transistors are upsized corresponding to the stack height. For example, for stack height of 16, all NMOS

transistors are sized 16 times larger than the unit sized transistors. Simulation results are shown in Fig. 8. It shows that stack height of 16 provides the worst delay, and stack height of 4 and 2 give similar delay. This indicates that higher stack height based design actually worsens the metrics, but one part that is missing in this simulation is the actual wire capacitance at internal node between each stage. To further investigate this, a wire capacitance of 50fF is added to each internal node. Simulation results in Fig.9 show that for stack height of 16, it achieves the best performance. The simulation above indicates that wire capacitance play an important role in determining the stack height to achieve the optimal performance. So accurate estimation of wire capacitance is necessary.

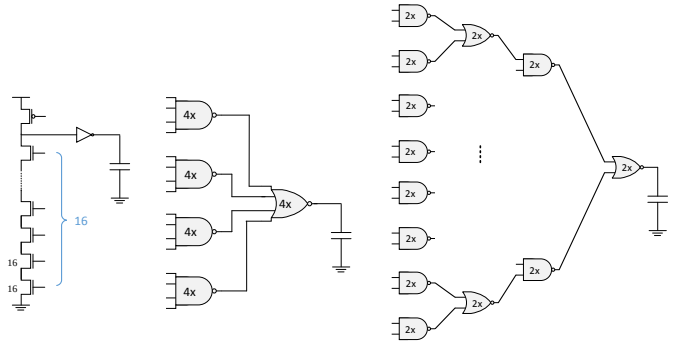


Fig.7. 16-input AND gate implemented with different stack height

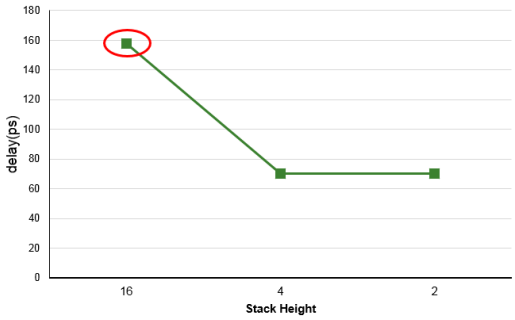


Fig.8. Simulation results without considering wire capacitance

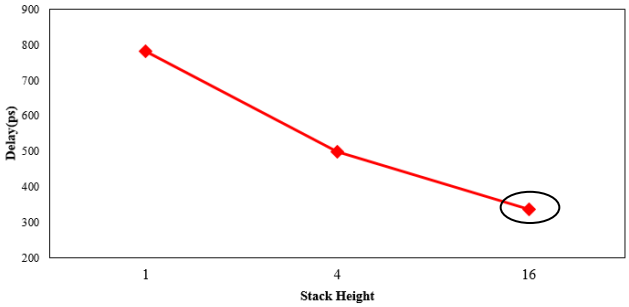


Fig.9. Simulation results with wire capacitance of 50fF

D. Estimation of Wire Capacitance

Wire capacitance usually includes two parts, one is intrinsic capacitance of the bottom and top metal plates, and another is the coupling capacitance from neighboring metal wires [12]. The total capacitance is the sum of the two. Normally, the coupling capacitance is much larger than the intrinsic capacitance. Based on the layout cells from the Free PDK15 [10], we estimated the wire capacitance in a wire capacitance calculator [13] developed by ASU. Fig. 10 shows the wire capacitance of a 5um-long metal1 wire. It shows that the realistic wire capacitance of the general logic cells could be in the order of several fF.

Dimensions	RLC
W = 0.050 um s = 0.050 um l = 5 um t = 0.20 um h = 0.20 um K = 2.2	R = 10.999 Ohm L = 0.004 nH M12 = 0.003 nH M13 = 0.002 nH M14 = 0.002 nH (k12 = 0.75 k13 = 0.5 k14 = 0.5) Cground = 0.04746 fF Ccouple = 0.56906 fF Ctotal = 1.23304 fF

Fig.10. Wire capacitance estimation

To show how wire capacitance will affect the performance of the 16-input AND cell with difference stack height, we sweep the wire capacitance from 0 to 50fF. Simulation results are shown in Fig. 11 and Fig. 12. Based on the wire capacitance estimation done above, it shows that stack height of 4 is preferred. If wire capacitance increases and exceeds 16fF, the stack height of 16 will achieve the best performance.

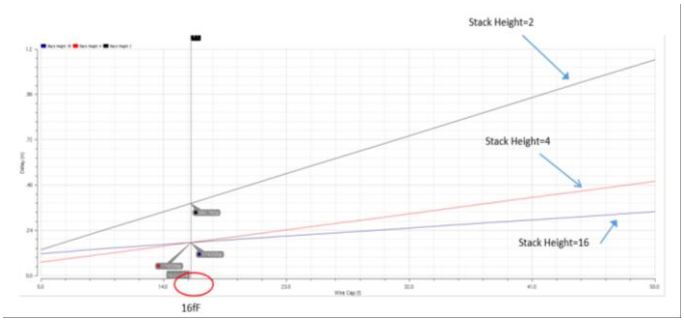


Fig. 11. Wire capacitance vs. Delay (from 5fF to 50fF)

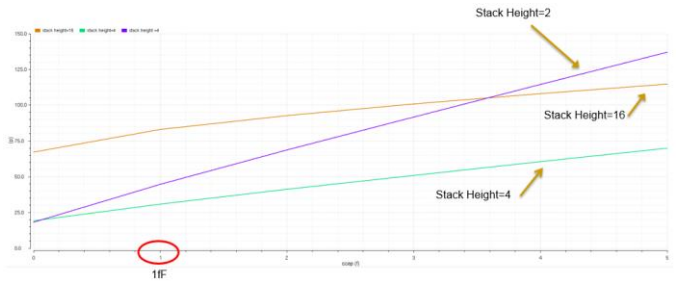


Fig. 12. Wire capacitance vs. Delay (from 0fF to 5fF)

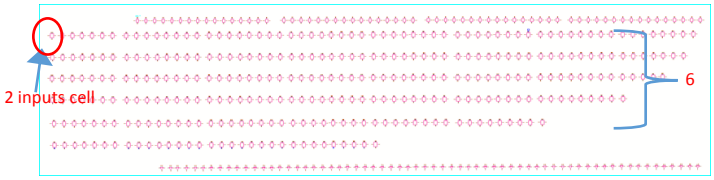
The analysis in this section suggests that accurate estimation of the interconnect capacitance is necessary when choosing the optimal stack height.

IV. Case Study: A 64-bit Kogge Stone Adder

Kogge-Stone adder [14] is a parallel prefix form of carry look-ahead adder that has a O(log n) time complexity. 64-bit adders have been widely used in modern processors and embedded systems [15]. A central idea of generating a faster Kogge-Stone adder is to use multi-fan-in (also called high-radix) within a Kogge-Stone prefix graph [16]. All of these made a 64-bit Kogge-Stone adder an ideal test vehicle for our stack height analysis that balances the parallelism and performance.

A. *Implementations*

Fig. 13 shows the two implementations of a 64-bit Kogge-Stone adder. (a) is radix-2 based and (b) is radix-8 based. It is clear that the radix-8 has fewer logic depth (6 compared to 2). The difference is that the internal operation (dot operation) blocks have different fan-in. For radix-8 implementation, each 8-input cell circled in Fig.13. (b) can be implemented in two ways. Shown in Fig. 14 are the two versions, the first implementation is to use one-stage fan-in of 8 gates, and another is to use 2-stage fan-in of 4 gates. The tradeoffs are between gate capacitance and stack height.



(a)

(b)

Fig.13. Schematic of 64-bit Kogge-Stone Adder with (a) Radix-2 (b) Radix-8

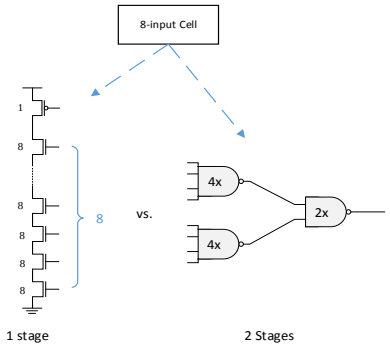


Fig.14. Two implementations of 8-input cell

Fig. 15 shows the simulation results in 16nm node, it shows that the 2 stage one has the best performance. While radix-8 with 1 stage has the worst delay. This is due to that the gate capacitance driven by the previous stage of the 1-stage implementation is the worst. So in this design, radix-8 with 2 stages is preferred. This is actually equal to radix-4 implementation. Simulations across 5 different technology nodes are shown in Fig. 16, and the performance improvement for the optimal implementation is shown n Fig. 17. It shows that at 16nm, the improvement can be ~6%.

B. *Area Estimation*

So far, we have demonstrated that the radix-8 implementation with 2 stages have the best performance. But since the size of each transistor in the stacked is increased correspondingly to ensure the balanced driving strength, the area of the higher stack height based design will be large. As an estimation based on the layout from the NanGate cell library [11], the 4-input NAND is 1.5 times larger than 2-input one, and each of the 8-input cell is about 5 times larger than the 2-input cell. As a result, the radix-8 implementation with 2 stages is 1.5 times larger than the radix-2 version. We believe that the area overhead can be eliminated by optimal physical design and floor plan.

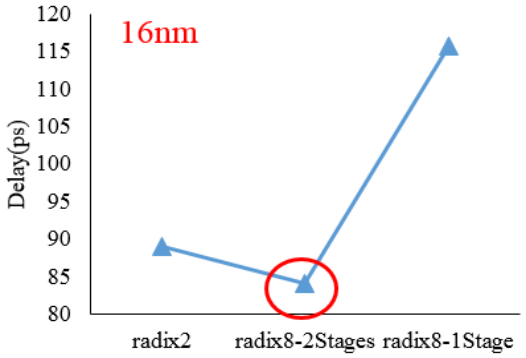


Fig.15. Simulation results with different implementations

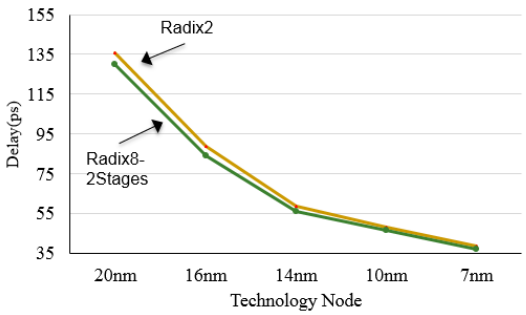


Fig.16. Simulation results across five technology nodes

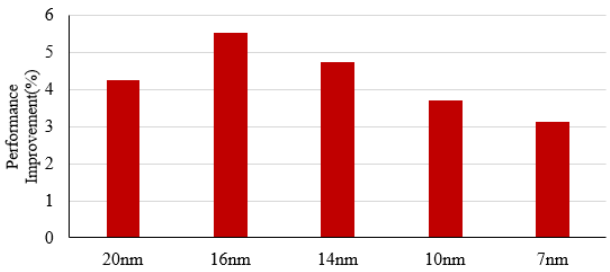


Fig.17. Performance Improvements across five technology nodes

V. DISCUSSIONS

A. *Tradeoffs*

Until now, we have analyzed how the stack height will affect the metrics of the FinFET circuit. It is clear that larger stack height will give optimal performance in some design cases, but not in all. So the tradeoffs need to be considered when stack height will be used as a design knob in FinFET circuit design.

First, as shown in Fig. 7, if the stack height is increased, each transistor that in stack needs to be upsized to balance the drive strength. The total size of the stack in terms of unit finger size is about *logH* times larger than the smaller stack one, where H is the ratio of the two stack height. As an example, Fig. 18 shows the number of unit finger transistor for

3 different versions of 16-input AND gate. It shows that the 16-stack height version uses the largest number of unit transistors. In terms of area, since FinFET based design is more compact and symmetric, some optimization techniques can be employed to reduce the actual area overhead for higher stack height.

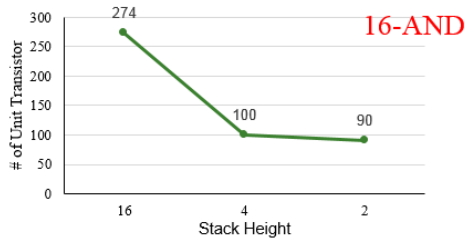


Fig.18. # of unit-finger transistors for three implementations

Second, as we increase the stack height and upsize the transistors, the gate capacitance increases. This will be translated to the performance penalty for the previous stage, there is an optimal point where the effect is minimal as shown in Section IV.

Last but not least, one big advantage of increasing the stack height is the reduction of leakage path. Shown in the Fig. 19 is an example of 16-NAND with stack height of 16 and 2. If we assume that the leakage with stack height of 16 design is 16I, where I is the leakage of the unit finger transistor, and the leakage of stack height of 2 design is (16+8+4+2)I, which is much larger. So higher stack has a potential of reducing static power consumption that is undesired especially in low power applications.

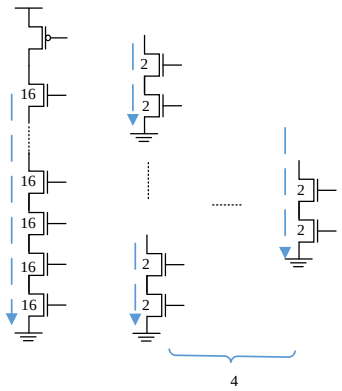


Fig.19. Leakage path reduction with higher stack height design

B. *Design flow for an optimal stack height*

As we have analyzed the tradeoffs when stack height is introduced as a design knob in FinFET circuit design, a design flow is proposed in this section.

Fig. 20 shows the proposed flow. In the first step, the stages to be stacked need to be picked. Theses stages could be either logic cells or certain gates within some cells. The next step is to estimate the interconnect capacitance between each stage for comparisons. It has been shown in Section III that wire capacitance has a big impact on determining the optimal stack height, so this step is crucial to the optimal design. By estimating the wire capacitance, some fair assumptions need to be made, metal layer and length for example. The capacitance values can be extracted from physical design or estimated based on accurate models, as did in this paper. Design metrics and targets need to be picked depending on the applications. For example, in ultra-low power design, minimizing leakage power is an important goal. Once the optimal stack height is picked, it is necessary to check if other design requirements are met. If it is met, the physical design could be done in the following step. During the physical design step, some layout optimization techniques could be employed. If some metric requirements are not met, the new balance need to pick, either through re-defining the stages or re-sizing certain cells. This flow could be applied in different phases of either top down design flow and custom flow.

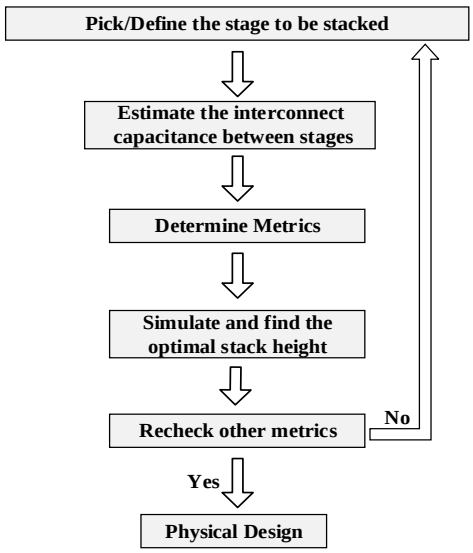


Fig.20. Design flow with stack height as a design knob

VI. Conclusion and Future Work

As FinFET technology is becomes promising, the uniqueness of the FinFET devices could be used as new design knobs for circuit designers. This paper introduces stack height as a new design knob in FinFET logic and circuit due to the insensitivity to body effect. Several logic gates and a 64-bit adder are simulated with PTM models. The results show the tradeoffs between the stack height and other metrics. It is clear that in some cases, higher stack height could improve the metrics a lot. A possible design flow is proposed to include the stack height into the whole design process. As future work, we will investigate more details into the physical aspects of the cells and include the accurate wire capacitance to the simulations. Also, stack height and sizing co-design techniques could be employed together to balance the tradeoffs between stack height, area and delay. More complex design will be explored in the future.

Acknowledgment

We'd like to acknowledge Prof. Mircea Stan for encouragement and guidance through the whole process. Also, we'd like appreciate the Free PDK group at NCSU for the open source FinFET design kit and PTM group at ASU for developing the FinFET simulation models.

REFERENCES

[1] P. Mishra, A. Muttreja, and N. K. Jha, "FinFET circuit design," Nanoelectronic Circuit Design, Springer, 2011.

[2] Poljak, Mirko, V. Jovanovic, and Tomislav Suligoj. "SOI vs. bulk FinFET: body doping and corner effects influence on device characteristics." *Electrotechnical Conference, 2008. MELECON 2008. The 14th IEEE Mediterranean*. 2008.

[3] Hook, T. B., et al. "SOI FinFET versus bulk FinFET for 10nm and below." *SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2014 IEEE*. IEEE, 2014.

[4] Y. Bin, C. Leland, S. Ahmed, W. Haihong, S. Bell, Y. Chih-Yuh, C. Tabery, H. Chau, X. Qi, K. Tsu-Jae, J. Bokor, H. Chenming, L. MingRen, and D. Kyser, "FinFET scaling to 10 nm gate length," *IEDM Tech. Digest*, pp. 251-254, 2002.

[5] A. Bansal, S. Mukhopadhyay, and K. Roy, "Device-optimization technique for robust and low-power FinFET SRAM design in nanoscale era," *IEEE Trans. Electron Devices*, vol. 54, pp. 1409-1419, Jun 2007.

[6] Andrade, M. G. C., et al. "Floating body effect on n-channel bulk FinFETs for memory application." *Devices, Circuits and Systems (ICCDCS), 2014 International Caribbean Conference on*. IEEE, 2014.

[7] Chiah, Siau Ben, and Xing Zhou. "Floating-Body Effect in Partially/Dynamically/Fully Depleted DG/SOI MOSFETs Based on Unified Regional Modeling of Surface and Body Potentials." *Electron Devices, IEEE Transactions on* 61.2 (2014): 333-341.

[8] Sinha, Saurabh, et al. "Exploring sub-20nm FinFET design with predictive technology models." *Proceedings of the 49th Annual Design Automation Conference. ACM*, 2012.

[9] Mahor, Vikas, and Manisha Pattanaik. "Highly robust Finfet based wide Fan-in dynamic OR gate with dynamic threshold voltage control." *Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on. IEEE*, 2014.

[10] FreePDK15: http://www.eda.ncsu.edu/wiki/FreePDK15:Contents

[11] NanGate: http://www.nangate.com/?page_id=2328

[12] Weste, Neil, et al. "CMOS VLSI Design: A Circuits and Systems Perspective."

[13] Interconnect Capacitance http://ptm.asu.edu/

[14] Kogge, Peter M., and Harold S. Stone. "A parallel algorithm for the efficient solution of a general class of recurrence equations." *Computers, IEEE Transactions on* 100.8 (1973): 786-793.

[15] Murthy, C. H., and T. Santhosh Kumar. "A Novel High Performance 64-bit MAC Unit with Modified Wallace Tree Multiplier." *Proceedings of ICETET* 29 (2014): 30th.

[16] Held, Stephan, and Sophie Theresa Spirkl. "Binary Adder Circuits of Asymptotically Minimum Depth, Linear Size, and Fan-Out Two." *arXiv preprint arXiv:1503.08659* (2015).